Б. А. Путко

Определение характера роста основной тенденции временного ряда при малом количестве наблюдений

Рассматривается задача определения характера роста короткого временного ряда (альтернативный выбор между линейным и экспоненциальным трендом). Предлагается простой критерий выбора. Результат может быть использован, в том числе, в медицине, в частности, при интерпретации динамики показателя онкомаркера.

Ключевые слова: линейный тренд, экспоненциальный тренд, МНК-оценка, гетероскедастичность, критическое значение, онкомаркер.

JEL classification: C22, C63.

1. Введение

Данная работа посвящена проблеме различия линейного и экспоненциального роста значений временного ряда при малом количестве наблюдений (4–10) и сравнительно небольшом увеличении наблюдаемых значений (порядка удвоения). Решение такой задачи в некоторых конкретных случаях помогает по наблюдаемым значениям определить природу процесса. Одним из важных примеров является установление предположительного диагноза на основании медицинских анализов, проводящихся периодически через равные промежутки времени (например, ежегодно). В частности, для выработки врачебной стратегии при росте показателей онкомаркера чрезвычайно важно различать экспоненциальный рост, характерный для онкозаболевания, и линейный рост, вызываемый, как правило, неспецифическими причинами. При этом именно удвоение (от середины референтного диапазона до его верхней границы) во многих случаях является сигналом для принятия решения.

2. Постановка задачи

Математическая формулировка задачи следующая: рассматривается модель

$$y_t = \alpha + \beta t + \gamma e^{\lambda t} + \varepsilon_t, \tag{1}$$

где ε_t — ошибки регрессии, причем $Corr(\varepsilon_t, \varepsilon_s) = 0$ при $t \neq s$ и $\sigma(\varepsilon_t) = \delta_t = pE(y_t)$, где p — постоянная величина (процент погрешности), т. е. модель является линейно гетероскедастичной. Будем также предполагать, что $E(\varepsilon_t) = 0$ и ошибки регрессии нормально распределены.

Случай линейного роста предполагает $\lambda = 0$. Для экспоненциального роста $\beta t << \gamma e^{\lambda t}$, на практике можно считать, что в этом случае $\beta = 0$.

Обратим внимание, что речь идет не о спецификации модели (Кремер, Путко, 2010; Магнус и др., 2007), а об альтернативном выборе.

Будем также считать, что имеется наблюдение y_0, предшествующее периоду роста и удовлетворяющее условиям: $E(y_0)=m=\alpha+\gamma$, $\sigma(y_0)=\delta_0=pm$.

Рассмотрим приращения $\Delta y_t = y_t - y_{t-1}$, $t=1,\dots,n$. Тогда

$$\Delta y_t = \beta + \gamma\left(1-e^{-\lambda}\right)e^{\lambda t} + \left(\varepsilon_t - \varepsilon_{t-1}\right). \tag{2}$$

Для определения характера роста y_t рассмотрим регрессию

$$\Delta y_t = c_0 + c_1 t + v_t. \tag{3}$$

В случае линейного роста $c_0=\beta,\ c_1=0$. При экспоненциальном росте $c_1>0$, причем МНК-оценка $\hat{c}_1$ параметра c_1 является смещенной. Это, а также малое количество наблюдений значительно затрудняет проверку условия $c_1=0$ с помощью доступного обобщенного метода наименьших квадратов (Айвазян, 2006; Магнус и др., 2007) — в модели (3) присутствуют и гетероскедастичность, и автокорреляция. К тому же процедура применения этого метода требует определенной квалификации специалистов и специфического программного обеспечения, что делает его практически невозможным для массового применения (например, в медицинских учреждениях). Требуется получить максимально простой критерий — типа сравнения двух чисел.

3. Критерий

Для решения этой задачи оценим модель (3) обычным методом наименьших квадратов. Непосредственным вычислением получаем следующее значение для оценки параметра c_1:

$$\hat{c}_1 = \frac{6}{n(n+1)}(y_n + y_0) - \frac{12}{n(n-1)(n+1)}\sum_{t=1}^{n-1} y_t. \tag{4}$$

Линейный рост. При этом

$$\hat{c}_1 = \frac{6}{n(n+1)}(\varepsilon_0 + \varepsilon_n) - \frac{12}{n(n+1)(n-1)}\sum_{t=1}^{n-1} \varepsilon_t,$$

$$E\left(\hat{c}_1\right)=0 \text{ (несмещенность)},$$

$$V\left(\hat{c}_1\right) = 4p^2\left[\frac{18}{n^2(n-1)}m^2 + \frac{18\beta m}{n(n-1)(n+1)} + \frac{6(2n-1)\beta^2}{n(n+1)^2(n-1)}\right]. \tag{5}$$

В случае удвоения показателя $y_n = 2y_0$ имеем $\beta = \dfrac{m}{n}$ и

$$\sigma\left(\hat{c}_1\right) = 2pm\sqrt{\frac{18}{n^2(n-1)} + \frac{18}{n^2(n+1)(n-1)} + \frac{6(2n-1)}{n^3(n+1)^2(n-1)}}. \tag{6}$$

Доверительный интервал имеет вид

$$\left(-2\sigma_{\text{lin}}\left(\hat{c}_1\right), 2\sigma_{\text{lin}}\left(\hat{c}_1\right)\right) \tag{7}$$

со значением доверительной вероятности $P_0 = 0.9545$.

Экспоненциальный рост. В этом случае непосредственное вычисление дает следующее значение:

$$\hat{c}_1 = \frac{6\gamma}{n(n+1)}\left[e^{n\lambda} + 1 - \frac{2}{n-1}\frac{e^{n\lambda} - e^{\lambda}}{e^{\lambda} - 1}\right] + \frac{6}{n(n+1)}(\varepsilon_0 + \varepsilon_n) - \frac{12}{n(n-1)(n+1)}\sum_{t=1}^{n-1}\varepsilon_i, \tag{8}$$

откуда

$$E\left(\hat{c}_1\right) = \frac{6\gamma}{n(n+1)}\left[e^{n\lambda} + 1 - \frac{2}{n-1}\frac{e^{n\lambda} - e^{\lambda}}{e^{\lambda} - 1}\right], \tag{9}$$

$$\sigma\left(\hat{c}_1\right) = \frac{6p}{n(n+1)}\sqrt{(\alpha+\gamma)^2 + (\alpha + n\beta + \gamma e^{n\lambda})^2 + \frac{4}{(n-1)^2}\sum_{t=1}^{n-1}(\alpha + \beta t + \gamma e^{t\lambda})^2}. \tag{10}$$

Будем предполагать, что $\beta = 0$ (сугубо экспоненциальный рост). Удвоение показателя $y_n = 2y_0$ означает условие $e^{n\lambda} = 2 + k$, где $k = \dfrac{\alpha}{\gamma}$:

$$E\left(\hat{c}_1\right) = \frac{6m}{n(n+1)(1+k)}\left[3 + k - \frac{2}{n-1}\frac{e^{n\lambda} - e^{\lambda}}{e^{\lambda} - 1}\right], \tag{11}$$

$$\sigma\left(\hat{c}_1\right) = \frac{6pm}{n(n+1)(1+k)}\sqrt{5(1+k)^2 + \frac{4k^2}{n-1} + \frac{8k\left[2+k-(2+k)^{\frac{1}{n}}\right]}{(n-1)^2\left[(2+k)^{\frac{1}{n}} - 1\right]} + \frac{4\left[(2+k)^2 - (2+k)^{\frac{2}{k}}\right]}{(n-1)^2\left[(2+k)^{\frac{2}{n}} - 1\right]}}. \tag{12}$$

Доверительный интервал имеет вид:

$$\left(E_{\exp}\left(\hat{c}_1\right) - 2\sigma_{\exp}\left(\hat{c}_1\right),\quad E_{\exp}\left(\hat{c}_1\right) + 2\sigma_{\exp}\left(\hat{c}_1\right)\right) \tag{13}$$

Заметим, что хотя интервал (13) не является доверительным для параметра c_1 (оценка $\hat{c}_1$ является смещенной), вероятность попадания в него наблюдаемого значения $\hat{c}_1$ равна доверительной вероятности.

Условием разделения линейного и экспоненциального роста является отсутствие пересечения интервалов (7) и (13).

Определим критическое значение $p_{\text{крит}}$ условием

$$E_{\exp}\left(\hat{c}_1\right) - 2\sigma_{\exp}\left(\hat{c}_1\right) = 2\sigma_{\text{lin}}\left(\hat{c}_1\right) = c_{\text{крит}}. \tag{14}$$

Это условие примыкания интервалов (7) и (13). Соответствующие величины $p_{крит}$ зависят от k и n. В таблице 1 приведены некоторые их значения.

Таблица 1. Критические значения $p_{крит}$

	n						
k	4	5	6	7	8	9	10
0.5	0.0129	0.0124	0.0119	0.0116	0.0111	0.0108	0.0102
1.0	0.0154	0.0148	0.0143	0.0138	0.0133	0.0129	0.0126
1.5	0.0175	0.0168	0.0162	0.0156	0.0151	0.0147	0.0142
2.0	0.0193	0.0185	0.0178	0.0172	0.0167	0.0161	0.0157

Как видно из табл. 1, величины $p_{крит}$ при различных значениях k и n находятся в диапазоне 0.01–0.02. При этом разделение линейного и экспоненциального роста происходит на уровне доверительной вероятности $P_0 = 0.9545$. На практике, как правило, приходится довольствоваться существенно меньшей вероятностью (порядка 0.8–0.9). В связи с этим предлагается использовать универсальное значение

$$p = 0.015\,. \tag{15}$$

Предположение (15) позволяет получать достаточно достоверные результаты при неизвестных значениях k (а они на практике и не бывают известными).

В таблице 2 указаны критические значения (14) в предположении (15).

Таблица 2. Критические значения $c_{крит}$

n	4	5	6	7	8	9	10
$c_{крит}$	0.0406m	0.0277m	0.0204m	0.0158m	0.0127m	0.0105m	0.0089m

В качестве оценки параметра m естественно взять само наблюдаемое значение y_0.

Критерий принятия решения следующий — предположение об экспоненциальном росте принимается, если наблюдаемое значение удовлетворяет условию $\tilde{c}_1 > c_{крит}$.

Экспериментальная проверка этого критерия проводилась автором с помощью компьютерного моделирования уравнения $y_t = \left(\alpha + \beta t + \gamma e^{\lambda t}\right)\left(1 + p\varepsilon_t\right)$, где ε_t — независимые стандартные нормальные случайные величины. Рассматривались случаи $\lambda = 0$ и $\lambda \neq 0$, $\beta = 0$. Параметры α, γ варьировались с помощью случайного генерирования. Практически во всех случаях следование сформулированному выше критерию приводило к верному выводу о характере роста.

Остановимся на некоторых практических вопросах.

На практике часто оказывается, что процентная погрешность достаточно велика — значительно больше значения $p_{крит}$. В частности, для медицинских анализов характерна погрешность порядка $p = 0.05$ (причем при отсутствии побочных факторов типа обострения воспаления). Для увеличения точности в качестве наблюдений можно взять средние значения нескольких замеров, выполненных с коротким промежутком. Так, усреднение четырех замеров позволяет увеличить точность в два раза. Значение $p = 0.025$ не разделяет характеры роста достоверно, однако на практике все же приводит, как правило, к верным результатам (компьютерное моделирование это подтверждает).

Таблица 1 показывает, что значения $p_{\text{крит}}$ уменьшаются с ростом n, т. е. при более быстром удвоении показателя достоверное разделение линейного и экспоненциального роста происходит при большей погрешности измерений. Различие это, однако, не слишком существенно, что и позволяет, по мнению автора, пользоваться универсальным значением $p = 0.015$.

Заметим также, что на практике, разумеется, никогда не наблюдается точное удвоение показателя, так что условие $e^{n\lambda} = 2 + k$ является идеализацией модели. Правильнее было бы рассматривать условие $e^{n\lambda} = 2 + \varepsilon + k$, где ε — относительно небольшая по модулю величина. Однако непосредственные вычисления показывают, что результаты изменяются незначительно, если ε принимает значения в нешироком диапазоне (примерно (– 0.2, +0.2)). Значение $p_{\text{крит}}$ при этом сохраняется, а величины $c_{\text{крит}}$ практически не изменяются.

4. Заключение

В заключение еще раз обратимся к медицинским приложениям полученных результатов, в частности, обработке данных динамики онкомаркера. В настоящее время, кроме самих его показателей, используется несколько критериев, рекомендованных для уточнения прогноза и используемых для принятия решения: выполнять ли биопсию — как правило, довольно травматическую процедуру, или пока оставить пациента под наблюдением (Матвеев и др., 1999; Пушкарь, 2003). Предложенный в данной работе критерий также может быть использован в этом ряду. По сравнению с общепринятыми методами, его использование технически более сложно и возможно лишь в специализированных учреждениях (например, в онкоцентрах, безусловно, обладающих статистическим программным обеспечением). В то же время, даже с учетом того, что в каждом индивидуальном случае предположения для применимости критерия могут нарушаться, он обладает большей достоверностью.

Автор благодарит Н. Ш. Кремера за полезные обсуждения и Н. И. Федорову за помощь при написании статьи.

Список литературы

Айвазян С. А. (2006). *Основы эконометрики*. М.: ЮНИТИ-ДАНА.

Кремер Н. Ш. (2007). *Теория вероятностей и математическая статистика*. М.: ЮНИТИ.

Кремер Н. Ш., Путко Б. А. (2010). *Эконометрика*. М.: ЮНИТИ.

Магнус Я. Р., Катышев П. К., Пересецкий А. А. (2007). *Эконометрика. Начальный курс*. 8-е изд. М.: Дело.

Матвеев Б. П., Бухаркин Б. В., Матвеев В. Б. (1999). *Рак предстательной железы.* Москва, РОНЦ.

Пушкарь Д. Ю. (2003). *Простат-специфический антиген и биопсия предстательной железы*. МГМСУ.